基于改进协同过滤算法的用户页面兴趣度预测研究 *

宋泊东，张立臣

(广东工业大学 计算机学院，广州 510006)

摘　要： 在海量的数据中发现用户的兴趣度是电子商务领域实现针对性信息推送的一种重要方法。根据大数据稀疏性特征，把奇异值分解方法引入协作过滤算法中进行互联网站点用户的页面兴趣度的计算和验证，提出了一种基于改进协作过滤算法的用户页面兴趣度预测算法。该算法可通过在网络日志文件中，提取显性用户评分数据存在的“虚假评分”，发现用户页面兴趣度和其影响因素。MATLAB 仿真结果显示：提出的基于改进协同过滤算法的用户页面兴趣度测量方法可有效克服海量数据的稀疏性，在预测准确性、测量速度方面都有很大提高。

关键词： 大数据；奇异值分解；页面兴趣度；协作过滤算法；数据稀疏性

中图分类号： TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.04.0282

Prediction of user page interest based on improved collaborative filtering algorithm

Song Bodong, Zhang Lichen

(*School of Computers Guangdong University of Technology*, *Guangzhou* 510006, *China*)

Abstract: discovering user interest degree in massive data is an important way to implement information push. This paper proposes a page interest prediction algorithm based on singular value decomposition and collaborative filtering. The algorithm can extract the "false score" in the dominant user score data, and find user page interest and influence factors. The results of MATLAB simulation show that the collaborative filtering algorithm based on singular value decomposition is accurate and efficient in predicting the interest degree of user pages under the unavoidable situation of massive data sparsity.

Key words: big data; singular value decomposition; page interest; collaborative filtering algorithm; data sparsity

0　引言

在海量数据中搜索发现用户兴趣度，是针对用户兴趣，实施提供个性化推荐的重要手段。目前，获取用户兴趣度（interest degree，ID）的主要有显式反馈与隐式反馈两种方式。不论是采用哪种方式，均通过采集计算用户访问页面次数、访问时间及页面操作动作来反映用户兴趣度。如夏义国等人[1]建立了一种神经网络模型，通过拟合用户访问网站页面次数和时间，分析计算用户兴趣度。李峰等人[2]建立一种基于隐式反馈的用户兴趣模型，通过分析用户的隐式反馈行为，计算、获取用户兴趣度。邢玲等人[3]构建了一种改进 K-means 算法的用户兴趣度衰减因子测算模型，计算分析用户的兴趣度。尹春晖等人[4]提出了一种基于用户浏览行为的用户兴趣获取算法，该算法通过分析用户页面浏览时间、速度和滚动条操作动作来解释用户兴趣度。王微微等人[5]提出了一种基于用户行为的兴趣度模型，分析用户的行为模式，结合用户的浏览内容，发现用户兴趣。于淼等人[6]提出了一种顽健的约束信息嵌入方法构建关系矩阵，降低了矩阵的稀疏性，进一步提高了算法的准确率。张艳梅等人[7]提出的适应用户兴趣变化的社会化标签推荐算法，陈臣[8]构建的图书馆小数据读者个性化兴趣预测与发现模型，李昌兵等人[9]对 Web 访问用户关联规则挖掘的研究，张朝恒等人[10]提出了基于社交网络信息的协同过滤推荐算法等关于页面兴趣度的算法研究。

上述的这些方法虽然采用不同方法和不同角度对用户兴趣度采集、测量进行了研究，但大都缺乏对大数据稀疏性的考量，所得结论往往和实际情况有较大差异。

因此，本文在前人研究基础上，考虑了海量数据稀疏性对用户兴趣度测量可能产生的影响，提出了一种改进协同过滤算法的网站访问用户兴趣度预测模型。该模型可以通过提取网络日志文件中显性用户评分数据，通过计算行与列中的任意两项之间的差值来预测用户对页面兴趣度和影响因素。

1　网络用户页面兴趣度计算和矩阵表示

1.1　页面兴趣度的计算

用户页面兴趣度是用户对网络页面感兴趣的程度。通过分析用户访问页面大小、访问次数、浏览时间、拉动滚动条、页

收稿日期： 2018-04-12；**修回日期：** 2018-05-23　　**基金项目：** 国家自然科学基金资助项目（61572142）；广东省自然科学基金资助项目（2015A030313490）

作者简介： 宋泊东（1993-），男，硕士研究生，主要研究方向为大数据、信息物理融合系统研究（1240279688@qq.com）；张立臣（1962-），男，教授，博士，主要研究方向为大数据、信息物理融合系统研究.

面保存等动作及页面发送字节数就可以揭示用户页面的兴趣度。

如将用户$i(U_i)$对页面$j(P_i)$的兴趣度设为$Interest_{i,j}$；将用户i对页面j 的访问次数设为n_{ij}；将用户i对页面j的平均浏览时间设为$t_{i,j}$；将页面j的大小设为$s_{b,j}$；将页面大小$s_{b,j}$对用户的浏览时间有影响的相关因素设为$T_{i,j}=\frac{t_{i,j}}{s_{b,j}}$；对页面兴趣度的次数影响因素设为$N_{i,j}=n_{i,j}$。但由于难以衡量$N_{i,j}$和$n_{i,j}$哪个影响因素更为重要，本文采用主成分分析方法就可获取这两个指标的权重，页面兴趣度的计算公式和标准化处理公式如下所示：

$$Interest_{i,j}=\alpha\times T_{i,j}+b\times N_{i,j}=\alpha\times\frac{t_{i,j}}{s_{b,j}}+b\times n_{ij} \quad (1)$$

$$T_{i,j}=\frac{T_{i,j}}{max(T_{i,j})} \quad (2)$$

$$T_{i,j}=\frac{N_{i,j}}{max(N_{i,j})} \quad (3)$$

1.2　用户页面兴趣度矩阵表示

将用户访问页面大小、访问次数、浏览时间、拉动滚动条、页面保存等动作及页面发送字节数进行前置处理后，转换为用户页面兴趣度 VSM 分析矩阵。本文采集回传的每一笔用户访问动作记录作为 1 份文件，并进行识别切割。形成不同的用户对不同页面的兴趣度 VSM 分析矩阵。如式（4）所示。

$$M_0=\begin{Bmatrix} i & \times & j\\ m & \times & n\\ \vdots & \vdots & \vdots \end{Bmatrix} \quad (4)$$

其中：M_0: 每个用户对不同页面的兴趣度；i:用户i;j:被访问页面次数；m表示用户的个数；n表示页面项的个数。

在得到用户页面兴趣度矩阵后，就可以对该矩阵的信息和某用户对某页面访问的操作信息进行标准化处理，然后采用协同过滤的方法来预测该用户对其他页面的兴趣度。

2　基于奇异值分解的 Slope One 算法改进与实证分析

2.1　算法改进思想

2005 年，Lemire 等人[11]提出的 Slope One 算法是通过计算群用户对一组项目的打分的平均差值的计算，来获得一个较合理的预测结果。黄义纯[12]提出的 Slope One 推荐算法改进，相较于传统的协同过滤推荐算法，Slope One 算法综合考虑了同组项目评价过的其他用户的打分情况和目标用户对其它项目的打分情况，所以更为准确和高效。Slope One 算法算法的用户评分矩阵示例如表 1 所示。

表 1　用户评分矩阵示例

用户	对项目 X 的评分	对项目 Y 的评分
A	3	4
B	2	4
C	4	?

表 1 为一个根据用户 A 和 B 评分矩阵，预测用户 C 对项目 X 的打分示例。

Slope One 算法的思想是：平均值可以代替某两个未知个体之间的打分差异。如 A 与 B 对项目 X 和 Y 的评分平均偏差是((3-4)+(2-4))/2=-1.5。即对项 X 的打分一般比项 Y 的打分要高 1.5。根据这种情况，采用 Slope one 算法就可以推测出用户 C 对项 Y 的打分是 4+1.5=5.5。根据上例，在已知用户评分矩阵的情况下，可以通过定义一个训练集 x，通过式（5）计算出项目 i 和项目 j 的平均差值$Dev_{j,\ i}$：

$$Dev_{j,i}=\sum_{u\in S_{j,i(x)}}\frac{u_j-u_i}{card(s_{j,i}(x))} \quad (5)$$

其中：$S_{j,i(x)}$为同时对项目 i 和 j 评价过的用户集合；u_i为用户 u 对项目 i 的评分；u_j为用户 u 对项目 j 的评分；$card(s_{j,i}(x))$为同时对项目 i 和 j 评价过的用户个数。

通过（5）计算$Dev_{j,\ i}$，就可以得到一个具有实时更新评分功能的对称矩阵。

当需要计算用户 u 对项目 j 的评价预测值$P(u)_j$时，通过训练集 x 中用户 u 已评价过的所有项目的集合 $S(u)$ 中与项目 j 有平均偏差$Dev_{j,\ i}$的集合R_i，就可计算出用户的兴趣度。具体计算公式如式（6）和（7）所示。

$$P(u)_j=\bar{U}+\frac{1}{card(R_j)}\sum_{i\in R_j}dev_{j,i} \quad (6)$$

$$R_j=\{i|i\in S(u),i\neq card(S_{j,i}(x))>0\} \quad (7)$$

其中：$\bar{U}$为训练集中用户 u 已评项目平均值；$card(R_j)$为训练集x中用户u已评价过的与项目j数量的平均偏差$dev_{j,\ i}$。

每个网站中都有大量页面，兴趣不同的每个用户往往仅仅访问自己感兴趣的页面。这种情况导致整个页面兴趣度矩阵呈现出稀疏性，使得寻找相似用户较为困难。这种情况也导致 Slope One 算法在计算较少用户关注的某个页面兴趣度数据时，无法预测新用户对该页面的兴趣度。因此，要解决这个问题，必须找出与用户页面兴趣度矩阵结构最相似的稠密矩阵做基础，进而采用 Slope One 协同过滤方法，过滤掉矩阵稀疏度，才能推导出用户对页面的兴趣度。

2.2　算法改进

奇异值分解(singular value decomposition， SVD)是一种应用广泛的机器学习算法。李卫疆等人[13]提出了融合信任传播和奇异值分解的社会化推荐算法，该算法采用特征分解处理评分矩阵。可有效避免数据拟合现象，克服零评分用户对相似度计算中的稀疏性、不精确问题，具有良好的推荐性能。

本文拟采用 SVD 算法，首先将采集的数据构建一个 m×n 阶矩阵 R，并分解为 $R=T_0S_0D_0,S_0=diag(\delta_1,\delta_{2,}\cdots\delta_r)$，计三个矩阵。其中，如果假设 $\delta_1\geq\delta_{2,}\cdots\delta_r\geq0$，则 T_0和D_0即可分别构成$m\times r$和$r\times r$的正交矩阵，即($T_0T_0=I,D_0D_0=I$)。式中的 r 是矩阵 R 的秩(r≤min(m, n))，那么，S_0则可以构成一个$r\times r$对角矩阵，将所有 r 值均设为大于 0 后，并按照大小降维排列后。就可以产生一个简化的近似矩阵。通常情况下，矩阵$R=T_0S_0D_0,T_0$，T_0、D_0和S_0都

必须为满秩，存在计算量大、运算复杂的缺陷。采用奇异值分解S_0，使之保留 k 个最大的奇异值，其他则用 0 来代替，就可得到含有 k 个奇异值的矩阵(k×r)。进一步删除S_0中值为 0 的行和列，就可得到对角矩阵 S ，同理，如果将矩阵T_0、D_0也采用降维方法简化为 T 和 D，就可得到矩阵$R_k = TSD, R_k \approx R$[12-14]。进而采用协同过滤算法，对用户兴趣度矩阵 R 进行规范化处理后，这样可以得到矩阵 R'。再用 SVD 算法就可得到输出矩阵 T、D 和描述用户在 k 维空间中的关系的$TS^{1/2}$（m×k）的用户矩阵及兴趣度大小计算$S^{\frac{1}{2}}D$（$n \times k$）项矩阵，引入协同过滤算法，计算$TS^{1/2}$（m×k）和$S^{\frac{1}{2}}D$（$n \times k$）就可得到用户页面兴趣度的大小。

1)推荐产生

将采集的数据标准化后的数据矩阵作为实现奇异值分解的基础矩阵$T_0T_0 = I, D_0D_0 = I$，然后采用向量空间方法计算$TS^{1/2}$的相似度。得到用户页面兴趣度的最近邻和相应的推荐集。

2)兴趣度计算

将矩阵$TS^{1/2}$和$S^{1/2}D$ 相乘，得到项内积，即可计算用户 u 对项 i 的兴趣度。测算公式如下：

$$I(u,i) = \bar{u} + TS^{1/2}(\mathrm{u}) \times S^{1/2}\mathrm{D}'(i) \tag{8}$$

其中，$\bar{u}$是用户u对已评分项的平均评分。

通过式（8）的计算，就可得到一个考虑新用户和新项目情况下无缺失值的用户评分矩阵，使用户的页面兴趣度矩阵形成100%密集，解决数据测算的稀疏性问题。在此矩阵上就可以计算得到用户对站点其他页面的兴趣度。

2.3 实证分析

2.3.1 兴趣度和稀疏度计算

本文使用淘宝网(https://www.taobao.com)的网络日志 2017 年 6 月份的数据进行分析。数据经过清理后，按上面所述方法进行处理后得到的页面兴趣度的公式如（9）所示。

$$\mathrm{Interesr}_{i,j} = 0.7071 \times \frac{t_{i,j}}{sb_j} + 0.8561 \times n_{i,j} \tag{9}$$

进而得出用户页面兴趣度矩阵M_0，其中用户 105 人，页面 168 个。部分用户页面兴趣度矩阵见表 2。

表 2　用户页面兴趣度矩阵M_0(部分)

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
U_1	0.307									
U_2	0.579		0.621							1.087
U_3	0.261			0.634						0.426
U_4	0.271	0.415		0.488	0.501		0.394	0.445		
U_5	0.277	1.432			0.33		0.445	0.595		
U_6	2.702	0.96	0.488	0.632	0.351	0.475	0.445	0.316	0.316	0.686
U_7	0.314	0.402	0.43		0.565	0.370				
U_8	2.072	0.65		0.768	0.337	0.436	0.445		0.331	
U_9	0.902	0.946	1.141	0.786				0.608		0.4
U_{10}	0.807			0.635						

矩阵M_0的稀疏度为 1- $\frac{827}{105\times168}$ =0.9468，这虽然说明数据存在极大的稀疏性，但仍在可接受的网络数据特性范围内。矩阵M_0标准化后变为M_{10}。从图 1 可以看出，随着矩阵稀疏度的不断增大，平均绝对误差 MAE 的值也随之不断增大。MAE 的定义如下：$r_{(u,\ i)}$，$r_{(\hat{u},\ i)}$，其中，$r_{(u,\ i)}$是对项目的预测评分，$r_{(\hat{u},\ i)}$是用户对项目的实际评分，而 MAE 的值即为 N 个此类评分对差的平均值，平均绝对误差 MAE 的值测算公式为

$$MAE = \frac{\sum_{u,i}\left|r_{(u,i)} - r_{(\hat{u},i)}\right|}{N} \tag{10}$$

如果得到的 MAE 的值越小，则对用户评价的预测就越准确，预测质量就越高。

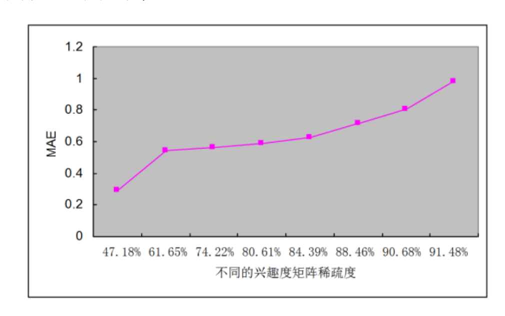


图 1　不同的页面兴趣度矩阵稀疏度对推荐算法的预测准确性的影响

本文通过 MATLAB 对奇异值矩阵M_1进行奇异值分解，使得$M_1 = T \times S \times D, DI$ ，矩阵M_1是一个 105×168 的矩阵，则通过奇异值分解得到的 T 为 105×105 阶矩阵， S 为 105×168 阶方阵，D 为 168×168 阶方阵。

2.3.2 5-fold 交叉验证

5-fold 交叉验证是一种为避免过度拟合而采用的边训练，边验证的机器学习算法。其原理是将训练数据随机分成 k 分，训练 k 次，直至找出最优解。为比较改进算法的有效性，本研究采用 5-fold 交叉验证方法对传统 Slope One 算法和改进的 Slope One 协同过滤算法（Slope One-after-SVD）进行比较，结果如图 2 所示。

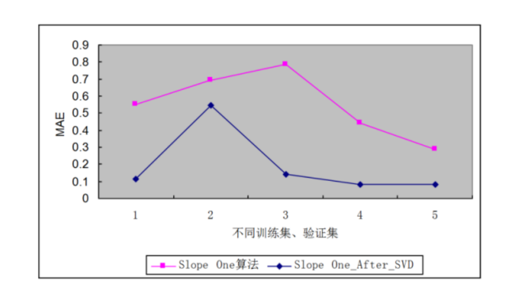


图 2　Slope One 和 Slope One_After_SV 算法

在 5-fold 交叉验证下的预测准确性比较，计算其平均值可以得到如式（11）的结果。

$$\overline{MAE_{Slopeone}} = 0.551, \overline{MAE_{SlopeoneAferSVD}} = 0.193 \tag{11}$$

通过平均值的计算和实验结果也可以看出，在数据稀疏性难以避免情况下，本文提出的基于改进协同过滤算法（Slope One-After-SVD）较之传统的 Slope One 算法在准确性方面有较大提高。通过训练集 3 和验证集 3 可以发现，仿真曲线呈现了截然相反的特征。这说明在处理数据稀疏度方面更具平滑特征。可有效避免因稀疏度未知或估计不足导致的验证失败。

3　结束语

本文根据大数据稀疏性特征，把奇异值分解方法引入协同过滤算法中进行互联网站点用户的页面兴趣度进行计算和验证。实验结果显示：在数据稀疏性难以避免情况下，本文提出的改进协同过滤算法的用户页面兴趣度预测方法，在不同的训练集、验证集上的预测准确性都有很大提高。但由于本文中的实验所采用的数据仅仅为一个特定的较小数据集，难以代表不同的和较大的数据集特性。因此，改进算法的适用性还需要在不同数据集、不同数据量下进行进一步验证。

参考文献：

[1] 夏义国，刘友华. 一种用户兴趣度计算与用户兴趣修正的改进方法 [J]. 计算机应用与软件, 2014, 34 (1): 46-55. (Xia Yiguo, Liu Youhua. An improved method of user interest calculation and user interest correction [J]. Computer Applications and Software, 2014, 34 (1): 46-55.)

[2] 李峰，裴军，游之洋. 基于隐式反馈的自适应用户兴趣模型 [J]. 计算机工程与应用, 2008, 44 (9): 76-79. (Li Feng, Pei Jun, You Zhiyang. Adaptive user interest model based on implicit feedback [J]. Computer Engineering and Application, 2008, 44 (9): 76-79.)

[3] 邢玲，宋章浩，马强. 基于混合行为兴趣度的用户兴趣模型 [J]. 计算机应用研究, 2016, 33 (3): 661-664, 668. (Xing Ling, Song Zhanghao, Ma Qiang. User interest model based on mixed behavior interest [J]. Application Research of Computers, 2016, 33 (3): 661-664, 668.)

[4] 尹春晖，邓伟. 基于用户浏览行为分析的用户兴趣获取 [J]. 计算机技术与发展, 2008, 18 (5): 37-39. (Yin Chunhui, Deng Wei. User interest acquisition based on user browsing behavior analysis [J]. Computer Technology and Development, 2008, 18 (5): 37-39.)

[5] 王微微，夏秀峰，李晓明. 一种基于用户行为的兴趣度模型 [J]. 计算机工程与应用, 2012, 48 (8): 148-151. (Wang Weiwei, Xia Xiufeng, Li Xiaoming. An interest model based on user behavior [J]. Computer engineering and Application, 2012, 48 (8): 148-151.)

[6] 于淼，杨武，王巍，等. 面向微博的多实体稀疏关系数据联合聚类 [J]. 通信学报, 2016, 37 (1): 151-159. (Yu Miao, Yang Wu, Wang Wei, *et al.* For micro-blog, multi entity sparse relation data joint clustering [J]. Communication Journal, 2016, 37 (1): 151-159.)

[7] 张艳梅，王璐. 适应用户兴趣变化的社会化标签推荐算法研究 [J]. 计算机工程, 2014, 40 (11): 318-321. (Zhang Yanmei, Wang Lu. A social tagging recommendation algorithm adapted to user interest changes [J]. Computer Engineering, 2014, 40 (11): 318-321.)

[8] 陈臣. 图书馆小数据读者个性化兴趣预测与发现模型的构建 [J]. 图书馆论坛, 2017, 37 (5): 98-105. (Chen Chen. Library small data reader personalized interest prediction and discovery model construction [J]. Library Forum, 2017, 37 (5): 98-105.)

[9] 李昌兵，凌永亮，汪尔晶. 基于兴趣度的 Web 访问用户关联规则挖掘 [J]. 计算机工程与设计, 2017, 38 (4): 852-856. (Li Changbing, Ling Yongliang, Wang Erjing. Web access user association rule mining based on interest degree [J]. Computer Engineering and Design, 2017, 38 (4): 852-856.)

[10] 张朝恒，何小卫，陈勇兵. 基于社交网络信息的协同过滤推荐算法 [J]. 计算机技术与发展, 2017, 27 (12): 28-34. (Zhang Zhaoheng, He Xiaowei, Chen Yongbing. Collaborative filtering recommendation algorithm based on social network information [J]. Computer Technology and Development, 2017, 27 (12): 28-34.)

[11] Lemire D, Maclachlan A, Slope One predictors for online rating-based collaborative filtering [J/OL]. (2008-09-15) . https://arxiv.org/pdf/cs/0702144. pdf.

[12] 黄义纯. Slope One 推荐算法改进 [J]. 现代计算机, 2017, 34 (35): 24-27. (Huang Yichun Slope One recommends algorithm to improve [J]. Modern Computer, 2017, 34 (35): 24-27.)

[13] 李卫疆，齐静，余正涛，等. 融合信任传播和奇异值分解的社会化推荐算法 [J]. 计算机工程, 2017, 43 (8): 236-242. (Li Weijiang, Qi Jing, Yu Zhengtao, *et al.* A social recommendation algorithm integrating trust propagation and singular value decomposition [J]. Computer Engineering, 2017, 43 (8): 236-242.)